Attorney Docket No. 81862.P182 *PATENT*

UNITED STATES PATENT APPLICATION
FOR

# GUARANTEED BANDWIDTH MECHANISM FOR A TERABIT MULTISERVICE SWITCH


First Named Inventor:

David Hughes


PREPARED BY:


BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CA 90025-1026
(408) 720-8300


**EXPRESS MAIL CERTIFICATE OF MAILING**

"Express Mail" mailing label number EL 617 210 061 US

Date of Deposit December 1, 2000

I hereby certify that I am causing this paper or fee to be deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Washington, D.C. 20231

Tina Domingo
(Typed or printed name of person mailing paper or fee)

12-1-2000
(Signature of person mailing paper or fee) Date

# GUARANTEED BANDWIDTH MECHANISM FOR A TERABIT MULTISERVICE SWITCH

## FIELD OF INVENTION

The present invention relates generally to communications systems and, more particularly, to traffic management in a network.

## BACKGROUND OF THE INVENTION

An asynchronous transfer mode (ATM) network is designed for transmitting digital information, such as data, video, and voice, at high speed, with low delay, over a telecommunications network. The ATM network includes a number of switching nodes coupled through communication links. In the ATM network, bandwidth capacity is allocated to fixed-sized units of information named "cells." The communication links transport the cells to a destination through the switching nodes. These communication links can support many virtual connections, also named channels, between the switching nodes. The virtual connections ensure the flow and delivery of information contained in the cells to the destination port.

However, if the switching nodes send a lot of traffic to a single destination port, the destination port may become congested. This local congestion of a single destination port may have an effect on the global traffic in the ATM network.

## SUMMARY OF THE INVENTION

A method including receiving control cells indicating that a destination port of an asynchronous transfer mode (ATM) network is congested, and reducing incoming traffic to the congested port to a guaranteed bandwidth of traffic until the destination port is uncongested, is disclosed.

Other features and advantages of the present invention will be apparent from the accompanying drawings and from the detailed description that follows.

## BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements, and in which:

**Figure 1** shows an embodiment of an ATM network.

**Figure 2** shows an embodiment of a line card.

**Figure 3** shows an embodiment of a method for scheduling a grant.

**Figure 4** shows an embodiment of a method for handling backpressure when a destination port is congested.

**Figure 5** shows a method for dynamically modifying allocation of bandwidth.

**Figure 6** shows an embodiment of a device for guaranteeing bandwidth.

## DETAILED DESCRIPTION

A terabit multi-service (TMS) switching platform includes a traffic manager (TM) for a fabric interface chip (FIC) to provide a method and apparatus for guaranteeing bandwidth in a terabit switching platform. In one embodiment, the method includes receiving control cells indicating that a destination port of an asynchronous transfer mode (ATM) network is congested, and reducing incoming traffic to the congested port to a guaranteed bandwidth of traffic until the destination port is uncongested. The TM can provide guaranteed bandwidth even if the line cards and switching fabric are from different vendors.

**Figure 1** shows an example of an embodiment of a TMS switching platform. Network 100 is a data transmission network with guaranteed bandwidth and quality of service. Data is transmitted through the network in cells. A cell is routed from its source to its destination through switching fabric 110, which contains switching elements (SE). Line cards 120 receive the cells and route the cells to an appropriate destination port through switching fabric 110. **Figure 2** shows an example of an embodiment of line card 120.

For example, as a cell passes from line card 120 to fabric 110, a queue engine 210 in the line card assigns the cell to a virtual output queue (VOQ) in VOQ device 220. The traffic manager (TM) 230 then schedules a departure time

for the cell. The cell then passes through the fabric interface (FIC) 240 to switching fabric and then to its destination port.

One function of the traffic manager 230 is to ensure a guaranteed bandwidth for unicast traffic on a port-to-port basis even when the destination port is congested. A congestion happens when multiple ports send more traffic than an optical carrier (OC) at the destination port can handle. For example, a data path from a traffic manager to an OC 192 destination port can only sink data at the OC 192 rate. If the incoming traffic exceeds this rate, then the destination port will be congested.

When globally informed of a congestion at a destination port, each TM in the network immediately rate-limits the traffic destined for the congested port to a guaranteed bandwidth. This guaranteed bandwidth may be pre-determined, using software for example. The guaranteed bandwidth is selected so that the total guaranteed bandwidth allocated to all egress ports for any given destination port does not exceed the rate of the destination port, which may be an OC 192 rate, for example. Otherwise, the oversubscribed traffic will consume the switching element (SE) buffer in the switching fabric, which may be shared among all destination ports. If this happens, it can have a global effect on traffic destined for non-congested ports.

Therefore, in one embodiment, the guaranteed traffic through a destination port is under-subscribed, so that a congestion can be quickly

relieved by the extra capacity of the destination port. The ingress traffic management functions of the TM provide this feature. The ingress traffic management functions in a TM manage the flow of unicast, multicast and control cells from a line card to a fabric interface chip (FIC). These functions of the TM may include grant scheduling, backpressure handling, bandwidth allocation, and speedup handling.

Grant Scheduling

The queue engine of a line card sends cell counts of its virtual output queues (VOQs) in its VOQ device to the TM in a round-robin fashion. The TM keeps track of these cell counts and uses them to issue grants to non-empty VOQs. The cell count may be carried in an 8-bit field in the request field, to allow a maximum cell count of 255 per VOQ. For example, the grant scheduler of the TM is busy for 8 superframe ticks (16 links x 8 cycles = 128 grants) before the next cell count update. The TM may issue more grants than the number of cells in a VOQ because of the handshake latency between the VOQ device and the TM. To handle this over-grant scenario, the line card simply drops the grants for empty VOQs and does not flag them as errors. In this embodiment, the grant scheduler may issue a new grant in four clock cycles. For example, given a clock frequency of 125Mhz, the TM grant scheduler can issue up to 31.25M grants/sec, which is 20% above the OC192 rate.

## Backpressure Handling

The TM receives backpressure information in different forms. An SE in the switching fabric broadcasts per port flow control cells to all ports when the cell count of a congested port exceeds a predefined threshold. The SE also broadcasts emergency flow control cells to all ports when its total buffer utilization exceeds a predefined threshold (the buffer may be shared among the destination ports in SE). The FIC can generate separate data and control link backpressure information for TM when one or more of the downstream SEs are congested.

## Bandwidth Allocation

The TM controls the cell flow from the queue engine by rate-limiting the grants issued to all VOQs. In one embodiment, there may be 122 VOQs, including 120 unicast VOQs, 1 multicast VOQ, and 1 control cell VOQ. **Figure 3** shows an embodiment of a method for the grant scheduler to pick a VOQ for the next grant. The current time (CT) is measured by a timer to represent the current time. There is theoretical departure time (TDT) counter and a intercell gap (ICG) counter for each unicast VOQ. The TDT counter defines the next cell's departure time and the intercell gap (ICG) counter defines how much a

TDT increments when updated. There may be no TDT and ICG counters associated with the multicast and control cell VOQ's.

When the status of a VOQ changes from empty to non-empty, step 310, its TDT is initialized with the current time in the CT, step 320. In the next cell tick, the CT increments by one and the TDT shifts to the left of CT. In each cell tick, the scheduler determines whether a TDT is less then CT, step 330. If so, the scheduler selects the smallest TDT that is less than CT, step 340, and issues a grant to the corresponding VOQ and re-calculates the new TDT based on "new TDT = current TDT + ICG," step 350. As CT increases with time, the new TDT will be served by the grant scheduler when it becomes the smallest TDT and its value is less than CT.

After the TDTs having values less than CT are served, the scheduler chooses among the VOQs which do not have any pending backpressure, step 360. At this stage, the VOQ selection may be based on either a round-robin method or a priority based method. In the round-robin method, each VOQ is treated equally. In the priority based method, the priorities among the unicast, multicast and control cell VOQs are programmable. However, the round-robin method may be maintained among the unicast VOQs even in the priority based method. In one embodiment, if the scheduler selects a VOQ which already has a future TDT time slot allocated, its TDT will be re-calculated based on the CT (i.e. new TDT = CT + ICG).

**Figure 4** shows an example of a method for handling backpressure when a destination port is congested. The TM receives control cells indicating that a destination port is congested, step 410. The incoming traffic to the congested port is reduced to the guaranteed bandwidth, step 420. For example, when either the unicast fifo (first-in, first-out) buffer or the control cell fifo buffer in the TM is filled up due to a data link backpressure from a FIC, the CT stops incrementing and the scheduler stops issuing grants to unicast VOQs in the current cell cycle. Stopping the CT implies the VOQs with guaranteed bandwidth do not gain credit when the data link backpressure is present. The credit may not be supported here because large catch-up traffic may be built up after the link backpressure is removed. The catch-up traffic then has to be treated as guaranteed traffic and can temporarily exceed the guaranteed rate, causing global congestion in the SE.

Multicasts cells received from the VOQ device are put in a multicast fifo buffer before they are duplicated. Copies of the first cell are sent to destination ports which are not backpressured. Head of line (HOL) blocking occurs if at least one destination of the first cell has backpressure. When the multicast fifo buffer in the TM is filled up due to link backpressure or port congestion, no grant is issued to the multicast VOQ in the current cell cycle. If the HOL blocking persists and is not caused by link backpressure, the TM drops the first cell after a timeout period and continuous with the next one.

In one embodiment, cells from the unicast fifo, multicast fifo and control cell fifo buffers may be sent to a FIC in a fixed priority: unicast (highest), control cell and multicast (lowest). The unicast may be given the highest priority because the unicast traffic may have a guaranteed bandwidth. The control cell VOQ may not have a guaranteed bandwidth, but may be traffic small, and can be easily accommodated by the speedup in the VOQ device and FIC interfaces. The multicast cells may be given the lowest priority, so that the additional bandwidth caused by cell duplication does not affect the unicast and control cell traffic. This prevents the multicast traffic from reducing the speedup advantage in the TM to FIC interface.

Speedup Handling

The speedup information about each VOQ is generated by the line card and is passed to the TM. A speedup flag is on when the throughput of its VOQ drops below a threshold. In the switching fabric, the scheduler may give the highest priority to traffic with speedup. The speedup turns off when the throughput of its VOQ is above a threshold.

In one embodiment, there is no need to speedup traffic because TM performs the bandwidth allocation in its scheduler. However, this raises another issue, fairness. For example, when TM grants 25% guaranteed bandwidth to a VOQ, it does not guarantee that 25% traffic will be constant bit

rate (CBR) traffic. CBR traffic supports a constant or guaranteed rate to transport services, such as video or voice, as well as circuit emulation, which require rigorous timing control and performance parameters. If the guaranteed traffic happens to be unspecified bit rate (UBR) traffic, it may be unfair to other VOQs which have non-guaranteed UBR traffic pending. Therefore, the TM should cut back its allocated bandwidth to a VOQ if the guaranteed traffic of that VOQ is not entirely CBR traffic.

One way to reduce this unfairness is to use the speedup information to dynamically modify the allocation of guaranteed bandwidth for a given VOQ, as shown in **Figure 5**. For example, when a speedup signal is off for a VOQ for a predetermined amount of time, step 510 the guaranteed bandwidth of that VOQ will be reduced by a small fixed amount, step 520. This slow downward adjustment continues until either the guaranteed bandwidth drops to zero, step 530 or the speedup signal turns on again, step 540. In this latter case, the guaranteed bandwidth will be incremented by a large fixed amount to ensure it satisfies the traffic demand quickly, step 550. This quick adjustment continues until either the original guaranteed bandwidth is reached step 570, or the speedup signal is off again, step 560.

**Figure 6** shows an example of a device for performing traffic management functions. Timer 610 measures the current time and is incremented by pulse device 615. A rate shaping circuit 620 is associated with a

corresponding VOQ. ICQ register 625 stores the intercell gap, which may be programmed using software. Adder 630 adds the ICG to the CT and outputs the TDT. Subtractor 640 subtracts the TDT from the CT. If the TDT is less than CT, then a valid signal is output from 645. Circuit 650 receives the valid signals from each rate shaping circuit and finds the VOQ having the smallest TDT. If there is no TDT less than CT, then having the comparator and round robin selector determines the next VOQ and next link to receive the next grant.

These and other embodiments of the present invention may be realized in accordance with these teachings and it should be evident that various modifications and changes may be made in these teachings without departing from the broader spirit and scope of the invention. The specification and drawings are, accordingly, to be regarded in an illustrative rather than restrictive sense and the invention measured only in terms of the claims.